

--	--	--

**Semester One 2017
Examination Period****Faculty of Business and Economics**

EXAM CODES: ETC1000 / ETW1000

TITLE OF PAPER: BUSINESS AND ECONOMIC STATISTICS – PAPER 1 OF 1

EXAM DURATION: 2 hours writing time

READING TIME: 10 minutes

THIS PAPER IS FOR STUDENTS STUDYING AT: (tick where applicable)

- | | | | | |
|------------------------------------|--|------------------------------------|--|--|
| ☐ Berwick | ☐ Clayton | ☐ Malaysia | ☐ Off Campus Learning | ☐ Open Learning |
| ☐ Caulfield | ☐ Gippsland | ☐ Peninsula | ☐ Monash Extension | ☐ Sth Africa |
| ☐ Parkville | ☐ Other (specify) | | | |

During an exam, you must not have in your possession any item/material that has not been authorised for your exam. This includes books, notes, paper, electronic device/s, mobile phone, smart watch/device, calculator, pencil case, or writing on any part of your body. Any authorised items are listed below. Items/materials on your desk, chair, in your clothing or otherwise on your person will be deemed to be in your possession.

No examination materials are to be removed from the room. This includes retaining, copying, memorising or noting down content of exam material for personal use or to share with any other person by any means following your exam.

Failure to comply with the above instructions, or attempting to cheat or cheating in an exam is a discipline offence under Part 7 of the Monash University (Council) Regulations.

AUTHORISED MATERIALS

OPEN BOOK	NO
CALCULATORS	NO
SPECIFICALLY PERMITTED ITEMS	NO

if yes, items permitted are:

Candidates must complete this section if required to write answers within this paper

STUDENT ID: _____

DESK NUMBER: _____

INSTRUCTIONS TO CANDIDATES:

Answer ALL questions in this examination paper.

Paper is out of 105 marks

This includes a 5-mark Question for Bonus marks at the end of Question 3

Where you are asked to perform calculations, you should write out the solution as an equation containing the appropriate numerical values from within the question. You do not need to calculate exact values in order to receive full marks for that part of the question.

Question 1 (24 marks)

In this question we will present a number of statements about data and statistics. In each case the statement represents a misunderstanding of the theory and ideas behind the methods. Explain carefully what is wrong with the statements and what they should have said.

- a. The confidence interval tells us that we are 95% sure the sample mean is between 17.6 and 25.4.
- b. The p-value is 0.43, so we conclude that the null hypothesis is true.
- c. The variance of $\bar{X}$ is $Var(\bar{X}) = \frac{\sigma^2}{n}$. So the bigger my sample size, the smaller the variance, and the less accurately I can estimate the mean.
- d. When we add these extra variables into our multiple regression model, the R^2 is bigger, so the model is better.
- e. If two variables are independent, that means the joint probability is zero.
- f. We want to know if one subject in the course is tougher than the other. A sample of results for each subject shows a sample mean of 74% for one, and 69% for the second subject. Clearly the second subject is tougher.

(each part is worth 4 marks)

Questions 2-4 are based on data collected in a survey of subsistence farming households in some rural areas of East Africa. This data was used in the Group project for this Unit.

Question 2: Crop Production (25 marks)

Poor, rural households rely on production of food crops for most of their food needs, and also to earn some income so they can buy more food and other essentials. In this question we investigate the total production of crops among these households.

- a. Here are descriptive statistics for the number of tonnes of crops produced per year by the households in the sample.

<i>Crop production (tonnes per year)</i>	
Mean	1.508955
Standard Error	0.038812
Median	1.098235
Mode	0.201875
Standard Deviation	1.606378
Sample Variance	2.580451
Kurtosis	41.98425
Skewness	4.976513
Range	23.17372
Minimum	0.000937
Maximum	23.17466
Sum	2584.839
Count	1713

- i. Compare the mean and median for this data. What does this tell us about the shape of the distribution of this data? Is this a surprising pattern? Based on how the mean and median are calculated, explain why they give such a different result in this case. (5 marks)
- ii. Interpret the standard deviation for this data. Give a precise (technical) interpretation of standard deviation, using the formula below as a guide to your explanation.

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

(4 marks)

- b. Using data from just some of the regions in the data set, we have calculated crop yield - production per hectare of land owned by the household. Below is output from calculating the correlation between crop yield per hectare and land size.

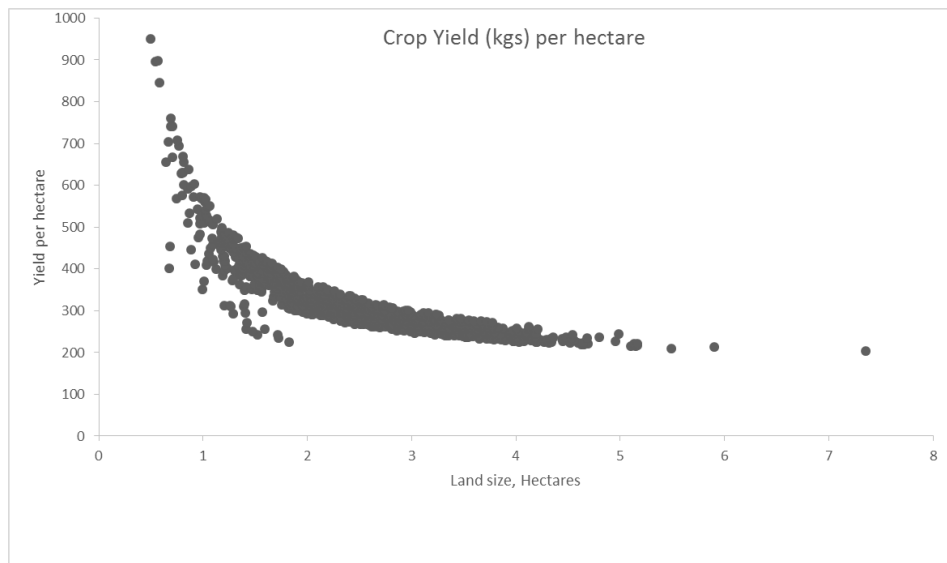
Dependent variable: Crop Yield / Standard Deviation of Crop Yield

Explanatory Variable: Land size / Standard Deviation of Land Size

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.792857							
R Square	0.628622							
Adjusted R Square	0.628239							
Standard Error	0.609722							
Observations	971							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	609.7631	609.7631	1640.2	1.217E-210			
Residual	969	360.2369	0.371761					
Total	970	970						
	Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	5.483255	0.054094	101.3659	0	5.37710049	5.589408951	5.343643699	5.62286574
X Variable 1	-0.79286	0.019577	-40.4994	1.2E-210	-0.8312749	-0.7544385	-0.843383265	-0.7423302

- i. Based on this output, what is the estimated correlation between land size and crop yield per hectare? What does this correlation estimate tell us about the strength and direction of the relationship between crop yield and land size?
(3 marks)
- ii. State the 95% confidence interval for the estimated correlation. Give a non-technical interpretation to this confidence interval.
(3 marks)
- iii. Note the output also shows a 99% confidence interval, which is wider than the 95% interval. Give an intuitive explanation for why this would be the case.
(3 marks)

To explore this relationship between crop yield and land size, we present a scatter plot for these two variables below.



- iv. The graph suggests that it was not appropriate to use the correlation measure in this case. Explain why this is so. (2 marks)
- v. Comment on this graph. What does it tell us about the relationship between yield per hectare and the amount of land a household owns? Can you suggest any practical reasons the relationship might follow this pattern? (5 marks)

Question 3 (29 marks): Food Shortages

Often there are times of the year when subsistence households experience food shortages (usually in non-harvest months). The survey asks households to recall whether they experienced food shortages and for how many months of the past year.

- a. In the following analysis we examine whether food shortages are related to the volume of crop production. The following variables are defined:

Experienced Food Shortage: =1 if the household experienced a food shortage sometime in the past year, 0 otherwise

Months Shortage: the number of months the household experienced food shortages in the past year.

Above average crop production: =1 if the household's crop production is above the average for all households, =0 if below average.

Now consider the following Excel Output, where the dependent variable is Number of months of food shortages.

SUMMARY OUTPUT						
Y=number of months food shortages						
Regression Statistics						
Multiple R	0.18172					
R Square	0.033022					
Adjusted R Square	0.032457					
Standard Error	3.509145					
Observations	1713					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	719.5163	719.5163	58.4303	3.49E-14	
Residual	1711	21069.42	12.3141			
Total	1712	21788.94				
Coefficients						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	5.582751	0.1198	46.6005	9.2E-307	5.34778	5.817721
above average crop production	-1.2962	0.169572	-7.64397	3.49E-14	-1.62879	-0.96361

- i. Interpret the two coefficients given in this output. What do they tell us about the relationship between crop production and food shortages?
(4 marks)
- ii. We would expect households with above average crop production to experience food shortages in fewer months of the year. Perform a hypothesis test to see if there is evidence to support this claim. Explain all the steps of your hypothesis test, and use the p-value approach.
(4 marks)

In the following output the dependent variable is “Experienced Food Shortages”.

SUMMARY OUTPUT						
Y = Experienced Food Shortages						
<i>Regression Statistics</i>						
Multiple R	0.166449					
R Square	0.027705					
Adjusted R Square	0.027137					
Standard Error	0.403601					
Observations	1713					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	7.941823	7.941823	48.75469	4.14E-12	
Residual	1711	278.7108	0.162894			
Total	1712	286.6527				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.855478	0.013779	62.08697	0	0.828453	0.882503
above average crop production	-0.13618	0.019503	-6.98246	4.14E-12	-0.17443	-0.09793

- iii. What parameter are we estimating with the coefficient of the intercept? (2 marks)
- iv. What do the values 0.828453 and 0.882503 tell us in this output? Explain carefully their interpretation. (3 marks)
- v. Note that both p-value are very small, and also that the 95% confidence intervals span a range that does not include zero. This is not a coincidence. Explain intuitively why this is the case. (5 marks)
- b. Food shortages are usually related to the timing of crop harvests. Using data collected over several years, analysts have come up with a model that identifies which months are more likely to produce shortages. Here is the estimated model:

$$\hat{y} = 0.43 - 0.022 \times \text{February} - 0.023 \times \text{March} - 0.011 \times \text{April} + 0.005 \times \text{May} + 0.015 \times \text{June} + 0.013 \times \text{July} + 0.020 \times \text{August} + 0.084 \times \text{September} + 0.143 \times \text{October} + 0.197 \times \text{November} + 0.082 \times \text{December}$$

where:

Y =1 if the household experienced food shortage in that month of the year, and 0 otherwise.

The month variables (February – December) are dummy variables equal to 1 if the value of Y refers to that month, and 0 otherwise.

- i. What is the meaning of the estimated intercept, 0.43? (2 marks)

- ii. According to the model, which month has the highest proportion of households with food shortages? What is the proportion in that month?
(2 marks)
- iii. Use the model to predict the proportion of households with a food shortage in June next year.
(2 marks)
- iv. **BONUS QUESTION:** Use the model to predict the proportion of households who experience a food shortage anytime in the year. N.B. It is not obvious how to do this – this is a bonus question for those who want to challenge themselves.
(5 marks)

Question 4 (27 marks): Children and Schooling

An indicator of well being of a household is the access that children have to schooling. This question will focus on the following variable:

$Y = 1$ if all school-aged children in the household attend school, $=0$ otherwise

- a. The following table shows the pivot table relating Y to the number of children in the household.

Number of Children in Household	All in School?		
	No	Yes	
1	4.2%	5.3%	9.6%
2	10.9%	9.2%	20.1%
3	15.9%	8.2%	24.1%
4	14.6%	6.9%	21.5%
5	9.4%	1.7%	11.1%
6	5.7%	1.2%	6.9%
7	2.9%	0.5%	3.4%
>7	2.9%	0.4%	3.3%
	66.6%	33.4%	100%

- i. Would you expect a relationship between the number of children in a household and their attendance at school? Briefly explain your reasoning. (2 marks)
- ii. If we chose a household at random, what is the probability it will have more than 6 children? (2 marks)
- iii. What percent of households have all children in school? (1 mark)
- iv. What is the probability a randomly chosen household will have 4 children and all school-aged children will be attending school? (1 mark)
- b. The following table shows the same information, but using a percentage of row format.

Number of children in household	All in School?		
	No	Yes	
1	44.3%	55.7%	100%
2	54.2%	45.8%	100%
3	66.1%	33.9%	100%
4	68.0%	32.0%	100%
5	84.3%	15.7%	100%
6	83.0%	17.0%	100%
7	85.3%	14.7%	100%
>7	87.9%	12.1%	100%
	66.6%	33.4%	100%

- i. The value 68% (4th row, 1st column) represents a conditional probability. Explain what the probability is, and explain how this value is calculated from the values in the previous table. (3 marks)

- ii. Note how the percentages in the “yes” column decline as the number of children in the household increases. Explain how this shows that the two variables are not independent. What does this tell us about the relationship between these two variables?

(3 marks)

- c. Next we estimate a multiple regression model that tries to identify the factors that make a household more or less likely to have all school-aged children attending school.

The model estimates are given below. The dependent variable is the 0/1 indicator for whether all school-aged children attend school. The explanatory variables are included in the regression output. The household head variables and the house roof variables all equal one if the statement is true, and zero otherwise.

SUMMARY OUTPUT						
Y = All Children in School						
<i>Regression Statistics</i>						
Multiple R	0.341783					
R Square	0.116816					
Adjusted R Square	0.112603					
Standard Error	0.462114					
Observations	1713					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	8	48.15848	6.01981	32.21634	1.84E-47	
Residual	1705	364.1019	0.213549			
Total	1713	412.2604				
<i>Coefficients</i>						
	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	
Intercept	0.385738	0.038096	10.12531	1.93E-23	0.311017	0.460458
Household head completed primary school	0.092034	0.024711	3.724414	0.000202	0.043567	0.140501
Household head completed secondary school	0.173965	0.040901	4.253323	2.22E-05	0.093744	0.254186
Number of working-age adults in household	0.028585	0.00999	2.861479	0.004268	0.008992	0.048178
Number of children in household	-0.07819	0.007329	-10.6692	9.04E-26	-0.09257	-0.06382
Number of people aged over 60 in household	0.009203	0.0098	0.939104	0.34781	-0.01002	0.028424
House has a new metal roof	0.187893	0.037906	4.956855	7.88E-07	0.113546	0.262239
House has an old (leaking) metal roof	0.127868	0.024396	5.241337	1.79E-07	0.080019	0.175718
House has thatched or other simple roof	0	0	65535	#NUM!	0	0

- i. Why do you think the last variable in the list has such strange values? What have we done wrong in the specification of the model?

(2 marks)

- ii. The research report discussing these results states that “comparing two households, where the first has one more child than the other, the first household has 7.8 percentage points lower likelihood of having all children in school”. What is wrong with this statement? What should it have said?

(2 marks)

- iii. The report also says the following: “Compare two households with the same number of adults, children and old people, and the same quality roof, but one has a household head who completed secondary school, and the other had a household head with no education. The household with better educated head will on average be 17.4 percentage points more likely to have all children in school.” What is wrong with this statement? What is the correct statement?
(2 marks)
- iv. Is there any evidence in the estimated model to suggest that having more people aged over 60 is good for children’s engagement with school? Explain your reasoning.
(4 marks)
- v. The R^2 in this model is 11.7%. How would you interpret this value. Does it suggest you have a good model here?
(2 marks)
- vi. Why do you think the type of roof a household has would be included in this model? What effect do you think this set of variables is seeking to capture in the model?
(3 marks)